RESEARCH PAPER

FAIR Science for Social Machines: Let's Share Metadata Knowlets in the Internet of FAIR Data and Services

Barend Mons[†]

Leiden University Medical Centre, Poortgebouw N-01, Rijnsburgerweg 10 2333 AA Leiden, The Netherlands

Keywords: FAIR science; Semantic publication; Metadata; Knowlets; FAIR principles

Citation: B. Mons. FAIR science for social machines: Let's share metadata Knowlets in the Internet of FAIR data and services. Data Intelligence 1(2019), 21-42. doi: 10.1162/dint_a_00002

Received: April 4, 2018; Revised: April 7, 2018; Accepted: May 3, 2018

ABSTRACT

In a world awash with fragmented data and tools, the notion of Open Science has been gaining a lot of momentum, but simultaneously, it caused a great deal of anxiety. Some of the anxiety may be related to crumbling kingdoms, but there are also very legitimate concerns, especially about the relative role of machines and algorithms as compared to humans and the combination of both (i.e., social machines). There are also grave concerns about the connotations of the term "open", but also regarding the unwanted side effects as well as the scalability of the approaches advocated by early adopters of new methodological developments. Many of these concerns are associated with mind-machine interaction and the critical role that computers are now playing in our day to day scientific practice. Here we address a number of these concerns and provide some possible solutions. FAIR (machine-actionable) data and services are obviously at the core of Open Science (or rather FAIR science). The scalable and transparent routing of data, tools and compute (to run the tools on) is a key central feature of the envisioned Internet of FAIR Data and Services (IFDS). Both the European Commission in its Declaration on the European Open Science Cloud, the G7, and the USA data commons have identified the need to ensure a solid and sustainable infrastructure for Open Science. Here we first define the term FAIR science as opposed to Open Science. In FAIR science, data and the associated tools are all Findable, Accessible under well defined conditions, Interoperable and Reusable, but not necessarily "open"; without restrictions and certainly not always "gratis". The ambiguous term "open" has already caused considerable confusion and also opt-out reactions from researchers and other data-intensive professionals who cannot make their data open for very good reasons, such as patient privacy or national security. Although Open Science is a definition for a way of working rather than explicitly requesting

[†] Corresponding author: Barend Mons (Email: barend.mons@go-fair.org; ORCID: 0000-0003-3934-0072).

for all data to be available in full Open Access, the connotation of openness of the data involved in Open Science is very strong. In FAIR science, data and the associated services to run all processes in the data stewardship cycle from design of experiment to capture to curation, processing, linking and analytics all have minimally FAIR metadata, which specify the conditions under which the actual underlying research objects are reusable, first for machines and then also for humans. This effectively means that—properly conducted—Open Science is part of FAIR science. However, FAIR science can also be done with partly closed, sensitive and proprietary data. As has been emphasized before, FAIR is not identical to "open". In FAIR/Open Science, data should be as open as possible and as closed as necessary. Where data are generated using public funding, the default will usually be that for the FAIR data resulting from the study the accessibility will be as high as possible, and that more restrictive access and licensing policies on these data will have to be explicitly justified and described. In all cases, however, even if the reuse is restricted, data and related services should be findable for their major uses, machines, which will make them also much better findable for human users. With a tendency to make good data stewardship the norm, a very significant new market for distributed data analytics and learning is opening and a plethora of tools and reusable data objects are being developed and released. These all need FAIR metadata to be routed to each other and to be effective.

1. INTRODUCTION

In 2005, I published the article *Which gene did you mean?* [1], which started with the sentence: "*Computational Biology needs computer-readable information records.*" The article got less than 40 citations over a decade and, in 2018, 13 years later, most data sets and metadata of other resource objects are still a "nightmare for machines", not only in biology, and we still hide our valuable data sets behind incomprehensible text, tables, figures and (frequently broken) links to supplementary data in classical articles. In the same article I complained about the communication of scientific results in text-only: "*Text mining? ....Why bury it first and then mine it again?*". An entire scientific sub-discipline has developed—and is still growing—attempting to recover machine-readable and unambiguous information from free text, with some impressive results, but still never recovering anywhere near the full extent of information hidden in the analyzed textual and graphical records. In 2006, Wilkinson and Good [2] stated the following:

> *The Semantic Web for the Life Sciences (SWLS), when realized, will dramatically improve our ability to conduct bioinformatics analyses using the vast and growing stores of Web-accessible resources. This ability will be achieved through the widespread acceptance and application of standards for naming, representing, describing and accessing biological information. Unfortunately, many key biomedical ontologies are hidden from the SW because they do not utilize resolvable URIs to name their components. Like un-hosted html pages, they are invisible and unusable without context-specific software because the ontology developers have focused on Semantic rather than on Web.*

So, we knew for many years that science would soon be overwhelmed by its own exploding ability to generate data. But, in hindsight, few colleagues picked up on the early warnings. This has led us into the current situation of data that are not findable, accessible and interoperable. Therefore, most data and the associated services and workflows cannot be reused.

Now that science rapidly becomes data-driven and machine-assisted, the wide and easy access to data and related services in formats that can be used by machines becomes really critical. In 2014 an international meeting conceptualized the FAIR guiding principles [3]. These principles emphasized the need for machine-actionability of research objects in modern science. In a follow-up publication, it was made clear that FAIR is not equivalent to open [4]. The term "open" as used in the catch phrases *Open* Access, *Open* Science and *Open* Source, apparently has many connotations in the minds of researchers and other citizens. Not only does it cause a lot of anxiety about privacy or security sensitive data to be opened up for everyone to see, and reuse, it also carries the association of "free" as in "gratis". It should be pointed out that the FAIR principles allow for data to be provided for reuse by the data owner under well-defined conditions, which means that highly sensitive data do not have to be open to participate in the FAIR data ecosystem. In the Open Source software domain, many different licenses are common practice, and we should optimally learn from the practices, successes and failures in that field [5]. Hybrid licenses of code still allow reuse of the code by others (and modifying it) but under defined conditions. Similar hybrid for data would not only allow for the combination of public and restricted data for research, but also for data owners and publishers to recover the significant costs associated with providing data for effective reuse for prolonged periods of time. Therefore, FAIR principles allow for *FAIR science*, a concept that fully encompasses the concept open science, but extends the concept and approach to include restricted data, provided that researchers have acquired permission to use the restricted data. However, the FAIR principles are exactly *principles*. The actual data and services implementations that support FAIR science should follow FAIR principles wherever possible, but choices will have to be made about terminology systems, persistent identifier (PID) systems, formats and many other aspects. The final goal, both for data as well as for all associated research objects (such as supplementary articles, software, workflows and compute) is *optimal reusability* by machines and humans, and frequently in *that sequence*. This article will mainly deal with the "Findability" aspect of FAIR and how that relates to the concept of machine actionable metadata.

2. THE INTERNET OF FAIR DATA AND SERVICES

In the recent report of the European Commission's High Level Expert Group for the European Open Science Cloud, we defined the need for a so-called "Internet of FAIR Data and Services" (IFDS) [6], referring to a virtual space where machines and people can find, access, interoperate and thus reuse each others' research outputs in a trusted, affordable and sustainable way. The IFDS should develop following the original hourglass model [7] (Figure 1), which underpins the successful and scalable growth of the Internet as we know it.

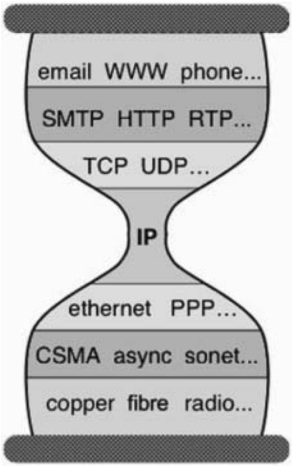


Figure 1. The hourglass model of the Internet architecture (for details see reference [6]).

Nothing in the envisioned Internet of FAIR Data and Services (IFDS) will likely be fully identical to the original Internet developments as the IFDS does not start in a greenfield and will build wherever possible on the current Internet infrastructure. However, there are clear similarities: In the classical hourglass layered systems architecture, the TCP/IP is usually placed in the narrow center of the hourglass, also referred to as the "spanning layer"[①]. In fact, all items below the spanning layer can be broadly classified as underlying network *infrastructure* and all levels above the narrow waist are leading to a wide variety of *applications*, with both sides having maximum freedom to make implementation choices. This is a basic principle to be followed in the IFDS as well: *Only set minimal necessary protocols and standards, and support a wide variety of implementation choices for data, tools and compute elements to participate in the growing IFDS* [6]. If we now try to translate the hourglass model to the IFDS, we deal with three distinguished, basic elements to be routed in order to find each other at the right time and place, and to be maximally used and reused. We have qualified these in the three broad categories DATA, TOOLS and COMPUTE. There are gray areas, because, for instance, software code (mainly covered under executable tools) can also be regarded as *data* and middle-ware could be classified as part of the compute infrastructure. We also realize that these boundaries may blur even further when data-driven and computationally assisted science will develop exponentially in the decades to come. However, for all practical purposes, we follow these practical broad definitions, and we basically want to treat all Digital Objects (DOs)[②] and the associated architecture in the IFDS according to the same principles. To ensure maximum findability off all digital objects, we here explicitly emphasize the need for sufficiently rich *machine-actionable metadata* such as what have been elaborated on in the FAIR principles and in several follow-up publications [4, 8]. Tools are defined mostly as software-type services that *act on data*, such as for instance virtual machines packaged to travel the IFDS

[①] See for a recent appraisal: Micah Beck, On the Hourglass Model. arXiv: 1607.07183 (https://arxiv.org/ftp/arxiv/papers/1607/1607.07183.pdf).

[②] https://www.internetsociety.org/resources/doc/2016/overview-of-the-digital-object-architecture-doa/

for distributed data analytics, but also, for instance, data repositories. Compute is defined as the actual compute processing elements that are needed for the tools to act on the data in a meaningful way. So, in order to keep the distinction between these three classes of "application layer" elements clear for the discussion, we argue here that we deal with three classes in three merging hourglasses, each with their own specific under-the-hood network and routing infrastructure. Obviously, wherever possible, generic elements of that infrastructure should be reused for all three elements. A way to express the IFDS elements (data, tools and compute) in relation to its underlying infrastructure to the hourglass image, we here propose the propeller image (Figure 2), acknowledging that from a purely architectural perspective we could still consider that a single hourglass.

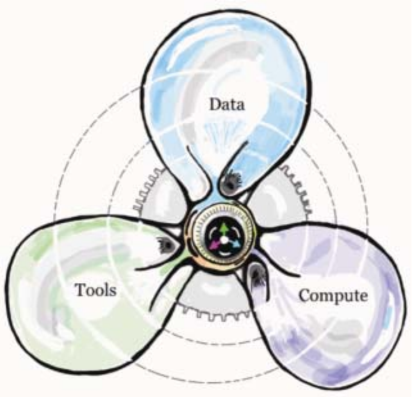


Figure 2. The merge of three hourglasses (data-infrastructure, tools-infrastructure and compute-infrastructure) into the image of a propeller with three blades and the underlying infrastructure. The narrow waist of the hourglass (minimal essential standards and protocols) is comparable to the center of this picture.

Intuitively, the IFDS would function most fluently in case the infrastructure (where possible the existing Internet infrastructure) would operate on a strong, common and globally interoperable networking and routing engine that could efficiently route data to tools, tools to data and both to the needed compute. These three elements will increasingly no longer reside in large centralized super storage and HPC facilities but will be more and more distributed all over the Internet. Therefore, additional performance aspects and security issues will have to be addressed but these are not the focus of this article and will be addressed separately. Here, we will mainly focus on the construction and the role of rich, machine readable and distributed metadata objects serving as a basis to locate, access and reuse the digital objects the metadata describe. For a more general description of metadata and the technology associated with them we recommend to read the information provided by the Center for Expanded Data Annotation and Retrieval[③].

A first very important aspect of our further reasoning is that we adopt the basics of the Digital Objects model and consider each digital object (from a single concept-reference, such as an identifier to a single machine-readable assertion to an entire database or software package) according to the following simplified scheme (Figure 3).

[③] CEDAR: https://metadatacenter.org

The first obvious prerequisite for the IFDS is that each digital object is assigned (and findable through) a unique, persistent and resolvable identifier (UPRI). The specific addition of the term *resolvable* here indicates the need to accept multiple UPRIs to point to the same concept, so they will correctly resolve to their defined meaning. There are several initiatives underway to repair the current undesirable situation where most data and services do not even fulfill this first criterion to participate in Open Science and the IFDS in general. We should build on these initiatives, and when they become community-adopted, we can follow them as well as contribute to their development wherever appropriate. For the sake of the argument in this article, we will assume that digital objects as *containers* have a UPRI.

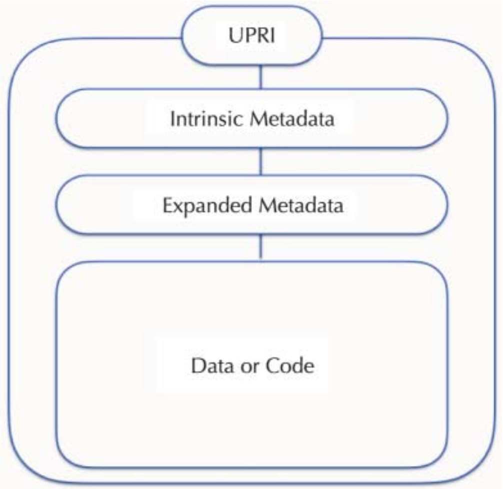


Figure 3. The simple Digital Object picture. The smallest conceivable Digital Object is a persistent identifier (PID) (a digital symbol referring to a particular concept). This concept could be an abstract unit of thought (in itself not a digital object) or it could refer to an actual digital object, such as another PID (could be a predicate or an object reference, but also an entire database). Each digital object that contains "information" should be adorned with metadata asserting things about the nature of that information. Here we distinguish, based on many discussions and the original DO architecture and in the context of the CEDAR platform between "intrinsic" metadata and "user defined" or "expanded" metadata, recognizing that sometimes the boundaries between those two may be rather arbitrary, we nevertheless believe that the distinction is practically meaningful. Typical intrinsic metadata describe the factual information that is "indisputable" about the digital object itself. For instance, assuming the digital object is a data set, the intrinsic metadata will describe the time of collection, the experiment they were part of, the creator, the equipment used to generate the raw data, the license, etc. However, in a world where Digital Objects (including research objects) will be increasingly and intensively reused by others than their creators, more subjective assertions about the digital object are also very important. These user-expanded metadata can be added by the original creators of the data, but may also be added by "reusers", and include subjective (and traceable/citable) assertions about errors, bias detected, etc. With the introduction of this second class of metadata, it becomes more and more important to also trace the provenance of the assertions made in the user defined metadata. Therefore, intrinsic metadata containers, expanded metadata containers and the actual containers holding the data elements or the core (in case of for instance a workflow) could also be treated as separate but permanently-linked digital objects, each with their own unique, persistent and resolvable identifier (UPRI) and thus form a stack of related metadata containers that contain (machine readable, FAIR) metadata of different nature, all asserting, however, relevant information about the data container.

However, in order to intelligently route data to tools, tools to data and both to compute (and in the future likely even mobile compute) we need more than just UPRIs for the containers. We need to describe the data or code containers with rich enough metadata in machine-readable format for both machine and humans (with lingual interface outputs and search capabilities for the latter) to Find, Access, Interoperate and thus effectively Reuse these components of the IFDS in a myriad of combinations in near real-time. As said, for each and every concept referred to in the metadata as well as, where possible, in the data themselves we need to enforce the use of UPRIs. Still, the choice for various UPRIs (even within the same domain) for the same concept is likely to persist at least for the foreseeable future and belongs to the first degree of freedom to operate away from the center of the hourglass. However, to enable this critical degree of freedom in the IFDS, which will be even more important when we really want to support interdisciplinary research and innovation, we need very high quality, robust and sustainable mapping services between UPRIs and human-readable terms that denote the same concept in digital objects. These mapping tables are critical infrastructure in the center of the propeller (Figure 2). A major problem is that currently, such services (for example BioPortal in the life sciences, OLS and FAIR Sharing) are built, maintained and funded largely by academic efforts and funded through volatile, few-year cycles of public funding, frequently even in fierce competition with "rocket science". A very important aspect of the IFDS will be to support the process of coordination within and across implementation, training and certification networks to minimize reinvention of redundant infrastructure components, including such things as thesauri and domain specific or generic ontologies protocols and other standards related elements of the IFDS. But, as said, we have learned that, classically, domains operate in silos and that even within domains multiple standards, vocabularies, languages and approaches will continue to emerge. This is not only a nuisance and a lack of coordination and discipline, it is also an intrinsic part of the creative process that should be supported in order to further our knowledge and drive innovation. This means that mapping tables, libraries to choose from, community standards registries, etc., will continue to be crucial elements of the IFDS support infrastructure.

Obviously, in the ideal IFDS, where machines form the majority of the first-line users, data, services and compute should all be machine actionable, and working seamlessly together, with human intervention being as minimally as possible, so that humans can focus on final interpretation and decision making based on patterns discovered by their machines. Not all research objects are digital (for example samples in biobanks) and not all digital data need to be entirely machine actionable to make the IFDS operational. However, every digital object, in order to participate, should have as a minimum FAIR metadata. Therefore, here we discuss the potential use of the existing Knowlet technology [9] to represent metadata of all objects as *concepts* in the IFDS, including data sets, workflows and compute facilities, without discussing the FAIR level of the actual underlying data, code, etc. So in principle, FAIR metadata can assert that the data they refer to is not (yet) FAIR. It should be emphasized here that the "Knowlet" concept is not prescriptive of the *format* in which its constituting components are expressed, but FAIR Knowlets should obviously be machine readable and preferably machine actionable like any other FAIR digital object.

3. THE STEPWISE BUILDING OF METADATA KNOWLETS

We will now build a stepwise argument to demonstrate how rich, machine readable metadata will lead us to a situation as described in Figure 4.

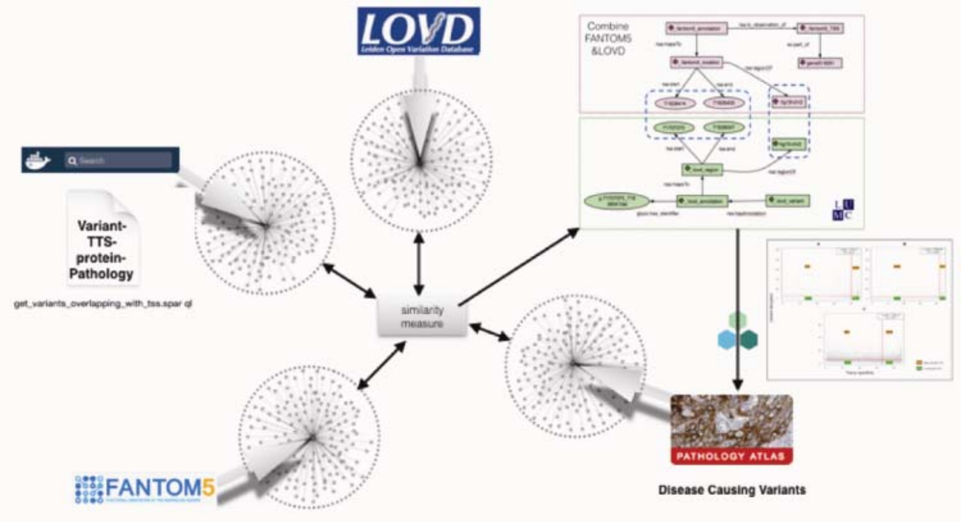
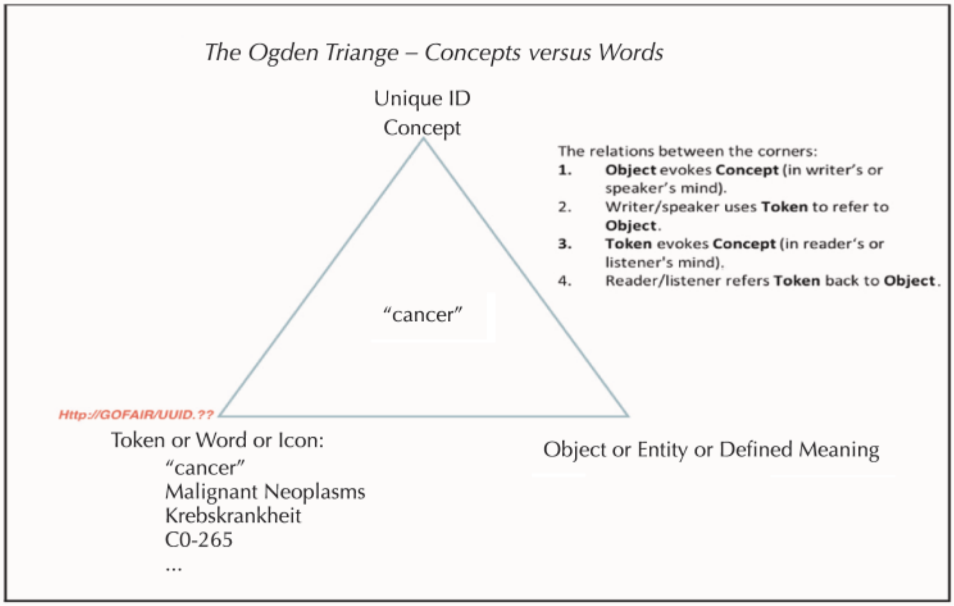

LOVD
Leiden Open Variation Database
Search
Variant-TTS-protein-Pathology
get_variants_overlapping_with_tss.sparql
similarity measure
Combine FANTOM5 &LOVD
FANTOM5
PATHOLOGY ATLAS
Disease Causing Variants

Figure 4. Assume that today, one would wish to ask the question: How many **genetic variants** that have been considered **potentially associated** with **disease** are not in **protein-coding sequences** (**genes**) but in **putative transcription start sites** and how many of the **proteins** associated with the **genes** related to these **transcription start sites** are **expressed** in **tissues affected** by the **disease** in question and is there evidence at the **protein visualization level** for **abnormal proteins** in **diseased tissue**? Imagine how many synonyms exist in literature for the **bold terms** in the query, as well as in free text metadata fields (if these are available at all) and consequently how long it would take a researcher to find all relevant databases to query, turn the data into a machine-processable format (as we would be dealing with well over 100,000 variants to test over all tissues and over 90,000 putative transcription start sites) and finally run a complicated machine actionable query like this? The figure shows how in the developing Internet of FAIR Data and Services, a linked-data-compliant query in a virtual machine format could automatically find the most relevant databases (in this case LOVD [10], FANTOM5[4] and the Human Protein Atlas[5]. Next, assuming that the data in the databases themselves are also FAIR, a machine would also be able to give some of the output shown in the figure in a fully automated fashion. Technically, this is already possible [11], but we would qualify most of the underlying infrastructure as "professorware" [12]. In the remainder of this article we will follow a step-by-step reasoning of how we could get to a scalable and ultimately industry grade version of such early implementations.

In 2009, Mons and Velterop [13] have defined concepts as the smallest, unambiguous unit of thought. This is in accordance with the Ogden or Semiotic triangle (Figure 5), where the concept is a unit of thought (irrespective of it being a real-world measurable entity), while there are many symbols that humans or machines may use to refer to that unit of thought and thus indirectly to the "actual instance" of that concept (as an entity or a common abstraction).

[4] https://www.nature.com/collections/jcxddjndxy

[5] https://www.proteinatlas.org/



The Ogden Triange – Concepts versus Words

The relations between the corners:

1. **Object** evokes **Concept** (in writer's or speaker's mind).
2. Writer/speaker uses **Token** to refer to **Object**.
3. **Token** evokes **Concept** (in reader's or listener's mind).
4. Reader/listener refers **Token** back to **Object**.

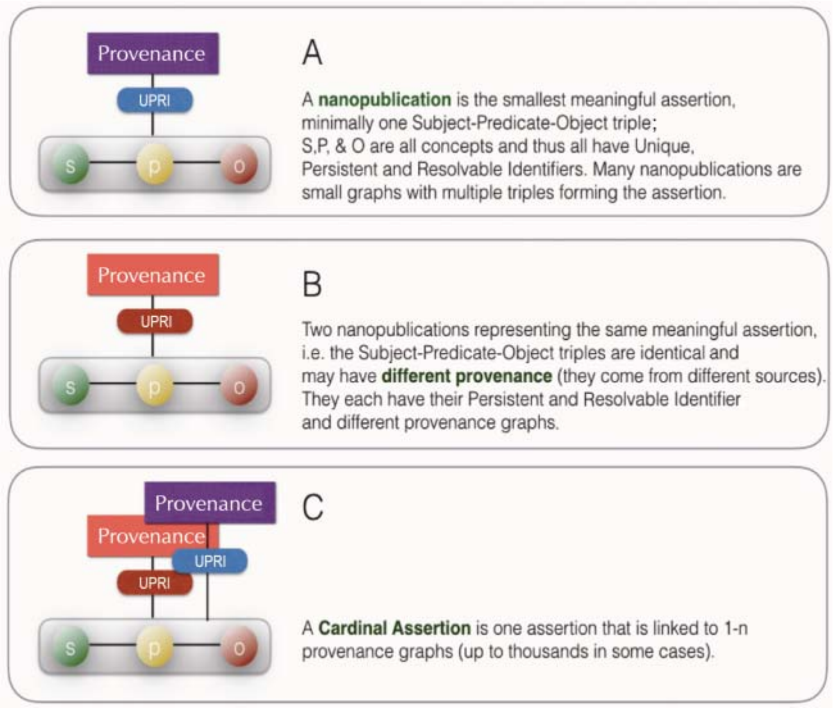
Figure 5. The semiotic triangle, based on the concept of cancer.

Groth, Gibson and Velterop [14] defined the anatomy of a nanopublication, a concept originally developed by Mons and Velterop [13]. In essence, a nanopublication is the smallest possible machine readable graph-like structure that represents a meaningful assertion (see Figure 6).

In the article *Calling in a million minds for community annotation* [9] we defined *Knowlets* in a broader sense and described their early use, but we first need to redefine Knowlets in the scope and context of the later definitions of concepts, nanopublications and cardinal assertions to place Knowlets and their proposed use in the line of logic of this article.

3.1 FAIR Knowlets

In current terminology, a Knowlet is a collection/cluster of Cardinal Assertions that share the same **subject**, and as such form a cluster of all cardinal assertions [15] that have been collected concerning (or about) that subject. The collective **objects** in the Knowlet constitute the "concept cloud" that has been directly associated with the subject of the Knowlet, and thus defines the subject according to all assertions about it collected so far (see Figure 7). Because a Knowlet is composed of individual cardinal assertions, and cardinal assertions are machine readable nanopublications, a properly constructed Knowlet is *FAIR "in principle"*. However, to be actually findable, search tooling and accessibility infrastructure is needed. Also, according to our earlier definitions of digital objects and concepts, a Knowlet is also a digital object, and it refers to a "concept" (which may or may not be a digital object in and of itself) and therefore each Knowlet needs a UPRI (Figure 8).



A **nanopublication** is the smallest meaningful assertion, minimally one Subject-Predicate-Object triple; S,P, & O are all concepts and thus all have Unique, Persistent and Resolvable Identifiers. Many nanopublications are small graphs with multiple triples forming the assertion.

Two nanopublications representing the same meaningful assertion, i.e. the Subject-Predicate-Object triples are identical and may have **different provenance** (they come from different sources). They each have their Persistent and Resolvable Identifier and different provenance graphs.

A **Cardinal Assertion** is one assertion that is linked to 1-n provenance graphs (up to thousands in some cases).

Figure 6. The single meaningful assertion in machine readable format is called a nanopublication. The smallest conceivable assertion has the structure of a subject, a predicate and an object. To form a nanopublication this "triple" needs to be published in machine readable format with full provenance and publication information (also in machine readable format) [14]. Note: Provenance and publication information is usually also in "triple" format (Panel A). The exact same assertion may appear in the Internet of FAIR Data and Services (IFDS) for multiple times (up to many thousands actually) and each of those identical assertions has a different provenance associated and thus by definition constitutes a unique nanopublication (Panel B). If we take the "cardinal" assertion that is common in all nanopublications asserting the same, we create a so-called Cardinal Assertion [15]. Cardinal Assertions are thus much less abundant than individual nanopublications in the IFDS. In principle, each Cardinal Assertion exists only once (as a unit of assertion) and it is "associated" with multiple, potentially many thousands of instances of nanopublications that assert the same, but differ in provenance.

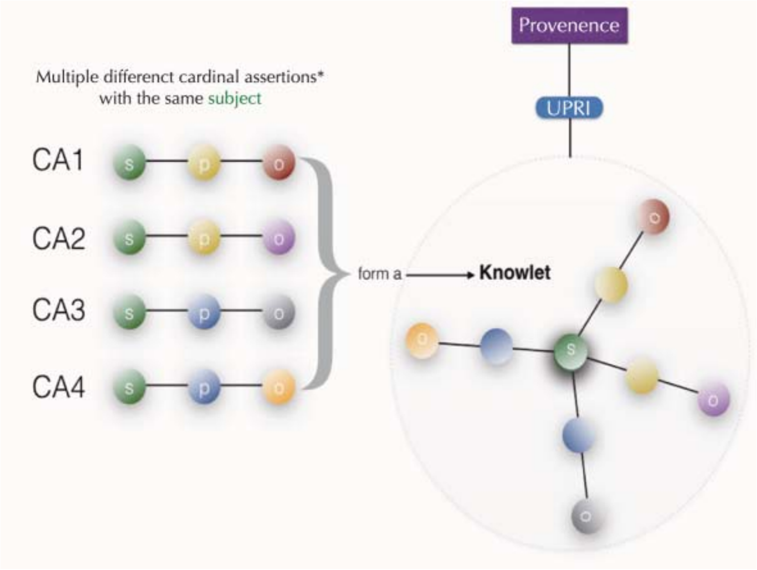


Figure 7. The Knowlet as a collection of cardinal assertions "about" a given subject. The objects effectively form the "conceptual context" of explicitly associated concepts. The predicates can range from very specific and explicit relationship descriptions such as "inhibits" or "is married to" to more generic and less explicit connections, such as "co-occurs in the same sentence as". Note: * UPRI and Provenance are not depicted for simplicity reasons.

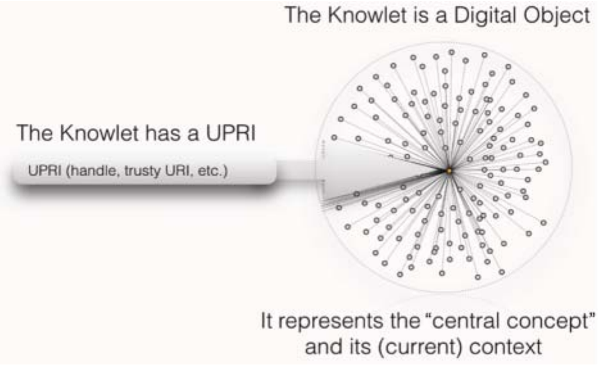

The Knowlet is a Digital Object
The Knowlet has a UPRI
UPRI (handle, trusty URI, etc.)
It represents the "central concept"
and its (current) context

Figure 8. The Knowlet is a digital object and needs to be findable, accessible, interoperable and reusable (i.e., FAIR) in its own right. It also may change over time, when more assertions are collected about the core concept. Therefore, each Knowlet in the Internet of FAIR Data and Services (IFDS) needs a unique, persistent and resolvable identifier (UPRI).

4. KNOWLETS AS METADATA

So far, we have used Knowlets mainly as a collection of cardinal assertions with the same subject, asserting something about abstract subject-concepts relations, such as genes, diseases, drugs, etc. These concept profiles or clusters can be used to determine "conceptual similarity" of their subjects in the knowledge space and to predict actual, meaningful associations between them, even if not explicitly asserted before[9,15]. However, Knowlets can *contextually define* any concept in the conceptual space of humans and machines, including a particular person, for instance, when composed of assertions such as [*ORCID*] [published][*Article doi*]. In such case, the Knowlet could define the "knowledge cloud" of the person with that ORCID, based on all relevant and specific concepts mentioned in all publications with that ORCID as an author. So, Knowlets could also define organizations, data sets, data bases, workflows and other services, virtual machines, etc., so in fact they can be used as metadata graphs for any *digital object* in the Internet of FAIR Data and Services. Taking this approach to its ultimate consequence, a Knowlet is a collection of cardinal assertions (each with full provenance) about the central subject and as such can be treated for the sake of argumentation in the context of this article as a form of machine readable and actionable metadata about the subject, regardless of whether the subject is an abstract concept such as a gene or love, or a real life entity, such as a particular person, a data set, a wearable device or a Dockerized Virtual Machine. It would therefore be needed to make the "status" of a Knowlet very clear, in a machine readable manner. In a way, we may even argue that any Knowlet is a form of "metadata about its core subject" as it effectively is a collection of assertions *about that subject*, regardless of whether the subject is an abstract concept or a real life entity, such as a data set or a workflow. We here propose to use Knowlets as a format to express and exchange machine readable (FAIR) metadata in the IFDS. It then follows that each concept in the conceptual space should have one (or multiple) Knowlet(s) as machine readable and actionable metadata that define that concept more precisely in a given context (see Figure 9).

5. TRUSTY URIS FOR KNOWLETS

Kuhn and Dumontier [16] defined the concept of *Trusty URIs*. They coined this approach originally for nanopublications and potentially larger data objects up to entire databases. They started with the aim to guarantee the integrity of nanopublications in terms of *immutability*. A trusty URI is in fact a form of a handle[6] where the final suffix is automatically created via hash algorithms and is based on a (selected part of) the content of the digital object it refers to. This means that if the content of that digital object changes, the hash code should also change, and therefore, the change is *detected*. In the case of a nanopublication or another relatively small and low-complexity digital object, in principle, all data elements or underlying URIs can be included in the hashing process. When we extend this approach to larger data sets, and more complex digital objects in general this might be impractical and also unnecessary. Also, it will not be practical to have a central, let alone manual, system to assign handles to digital objects such as assigning DOIs to articles for a fee, a critically useful and reliable service of Crossref[7]. As we will have many trillions

[6] https://www.internetsociety.org/resources/doc/2016/overview-of-the-digital-object-architecture-doa/

[7] https://www.crossref.org/

of nanopublications we would need to use mostly automatically generated handles. For example, of the FANTOM5 nanopublication data set (containing over 16 million individual nanopublications), each describing the genomic location of a putative transcription start site) [17] and it would be impossible to give each of those nanopublications a separate DOI for the costs incurred for an article DOI. It should be pointed out that, in principle, nanopublications are snapshots of assertions by a certain asserting person or machine at a given time and should as such be immutable. The same is explicitly *not true* for cardinal assertions. Here, the number of supporting nanopublications could increase, and there could be a contest of the validity of the assertion, and thus the provenance graph (in fact already a form of metadata) could change, thereby changing the trusty URI of the cardinal assertion. As Knowlets are in turn collections of cardinal assertions, the trusty URI is in this case not at all designed or used to *protect* it from being changed without detection. On the contrary, trusted URIs would change over time and *tracing that chain of Trusted URIs* as associated with the changing digital object would enable automatic tracking of the Knowlet over time in a blockchain-type fashion. If that Knowlet is in fact a metadata-graph pointing to a particular digital object (thus forming part of its metadata), automated tracking over time of changing digital object would become an intrinsic feature of digital objects in the IFDS. This would include the automated reporting of changes in, for instance, user defined metadata.

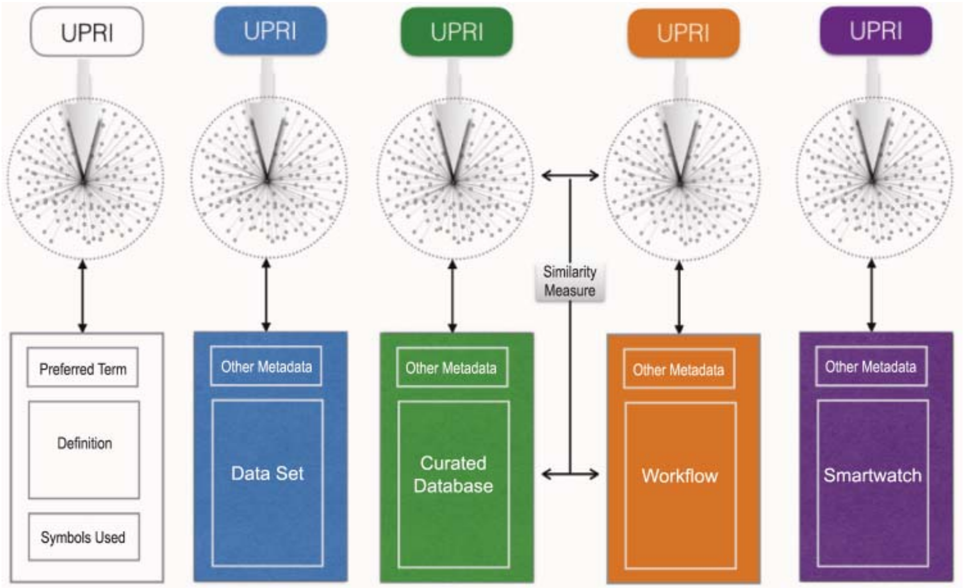


Figure 9. The Knowlet can be seen as a metadata container for the concept it represents. It can represent many different things from plain concepts like a gene or a person (ORCID record), to a data set, a data base, a work flow or any other thing in the Internet of Things.

In order to allow proximity matching of data objects beyond 1:1 conceptual overlap, we need to apply vector matching type techniques on metadata files, much as we do with biological concepts [18]. So, here we elaborate on a subtly different view on metadata than just looking at them as structured files describing the data set or service they refer to.

First, we generally define metadata in the context of the IFDS as assertions about digital objects. Second, we define metadata about a given digital object as assertions about that entity as a subject of the assertion.

These two steps may need some explanation, referring back to the earlier definition of a Knowlet. When we make a minimal unique assertion about a subject, we add a predicate (of a given class) and an object. The subject may be a certain abstract concept like a disease, a gene, etc., but it may also be a physical object, ranging from a data set (about which we make an assertion), a workflow, or a physical object in the IoT. As we argued in [13] we can define each unit of thought as a concept, so not only abstract concept identifiers in ontologies referring to genes, people or institutions effectively refer to concepts, but also identifiers that refer to data sets, databases, workflows and compute units can be seen as referring to concepts.

So, for all practical purposes we stick with the definition that each subject about which we can think and thus talk (or more precisely make assertions) is a concept, and we can define a Knowlet as the collection of all cardinal assertions made about a given concept at a given time. A Knowlet is composed of cardinal assertions, which are technically nanopublications in their own right and thus are assertions with provenance, supported by one or more nanopublications making the same statement. As argued in several other publications, nanopublications are assertions made by a certain source at a given time and context and thus are snapshots and are intrinsically immutable in nature and as argued in [16], in the case of nanopublications, trusty URIs have the role of safeguarding that immutability. However, cardinal assertions will not necessarily have an immutable character. The assertion as such is immutable, but its provenance changes, for instance when the number of nanopublications re-asserting the same assertion increases. So, cardinal assertions have versions. The same therefore holds for collections of cardinal assertions, such as Knowlets. Knowlets can become richer (more assertions about the central subject emerge), technically also poorer (when certain cardinal assertions are removed or suppressed for particular purposes). It follows that both cardinal assertions as well as Knowlets are intrinsically mutable and in this case the trusty URI serves as the track record of those changes, rather than as a way to prevent or just detect them. In this case the trusty URI comes close to an element of a blockchain-type infrastructure where the changes of, and the differences between Knowlets can be traced. By now treating each Knowlet as a metadata file, we have two elements of metadata files that were not there before.

First, each Knowlet (even if it is a new version of a previous Knowlet) has a unique, trusty URI. So no metadata file can even get coincidentally the same URI or PID. Consequently, a properly instructed machine can infer, that, even if 99.999% of the predicates and the objects in two Knowlets are identical, if they each—as a digital object—have distinct trusty URIs, they *represent different concepts*. Obviously, for instance, a new version of a database, is in fact a new concept but we should also be able to recognize its "near similarity" to the previous version of the same database. The actual conceptual similarity of Knowlets can be measured in multiple dimensions and many publications and practical (also commercial) applications already describe and use vector matching technologies to calculate the conceptual proximity of each pair of Knowlets [19]. So now the computer can deal with near-sameness or rather conceptual proximity of each pair of Knowlets in near real time in a dynamic concept space, while always being able to distinguish them as separate concepts based on their trusty URI. This already allows a form of high performance associative reasoning in computational environments and graph databases where the machine can deal with close proximity and hypothesis about implicit associations between concepts without those relations ever being made (ontologically) explicit.

Very important for our reasoning here is that, in the context of this article, we consistently think about Knowlets as metadata elements. In their current use (for instance in Euretos)[8] Knowlets describe a concept of the category gene, disease, etc. But even in this case we can see the assertions in the Knowlet of a given concept as assertions about that concept and thus they are in this definition context a form of metadata. And we know that a concept in the minds of people is largely defined in any case by the concepts they directly associate with it via a predicate (i.e., the Knowlet). Now, one step beyond this thinking, and faithful to treating a data set, a database, a workflow and an actual device in the IoT as a concept as well, the Knowlet representation of its metadata (all assertions made about the data set) should have a trusty URI (as a separate digital object), while the central trusty URI in that Knowlet (the subject) is represented by the URI referring to the data set, or workflow the Knowlets "talk about". Projects like CEDAR, ORCID and VIVO could build libraries of typical assertions people make about the semantic types they cover (respectively data sets, authors/contributors and institutions), and people can create Knowlets by filling standard metadata templates without being even aware of them.

6. ADDED VALUE OF ALL THIS FOR THE IFDS?

One of the first challenges in the IFDS is to find and locate the digital objects that need to be combined to run a particular data analytics job. FAIR metadata files, indexed by a variety of search (Google type) and fuzzy matching (Euretos-type) engines would power a meaningful combination of distributed data elements, with the services that can extract and discover meaning from them (see Figure 10). Of course, FAIR data points (FDP, defined here as any discoverable container with FAIR metadata) should be indexable by such services and the metadata should allow a targeted result, pointing to the assets that might be combined to discover new knowledge.

6.1 FAIR Data Points for Metadata

A properly deployed FAIR data point (FDP), with its features and containing rich metadata, supports many of the requirements for F, A, I and R. However, in order for the actual research objects hosted in a FAIR point themselves to be really *found* and *located* by machines their metadata should be indexed by a search or matching engine and of course these search engines need to know the existence of the FDPs. Two approaches can be used to achieve that. First, the people responsible for the FDP that contains the metadata of a digital object "somewhere on the Internet" should register the FDP in the search engine which will then index its metadata content. This mechanism has been already implemented in early FDP's and

[8] https://www.euretos.com/

prototypic search engines[9]. The FDP should preferably announce its existence to interested parties. Theoretically, the easiest way would be to broadcast the FDP's announcement. However, in the current network infrastructure (TCP/IP), for performance reasons broadcast is only allowed in local networks, not in the open Internet. The alternative is to multi-cast. While broadcast is a one-to-all, multi-cast is a one-to-many. The principle of multi-casting is that listeners (parties interested in receiving notifications from a given host) subscribe them in a list and, when the host has a message, it sends the message to all recipients registered in the list. The Internet protocols support multi-cast with the IP multi-cast (IGMP on IPv4 or MLD on IPv6). Here we could rely on the IP infrastructure and/or create a service that has a unique and immutable (trusty) URL and every FDP or FDP-compatible application in the world can register itself and notify whenever there is an update. It is proposed that we engage with the Internet Assigned Numbers Authority (IANA) to register a service (the FDP metadata notification service) and a transport port number (like the port 80 for HTTP, 21 for FTP, etc). A trusted, not for profit entity should host and maintain this registration service.

As a next step, changes in metadata FDPs or "Knowlets" should be detectable and traceable for search and matching services. This could be achieved by announcing changing trusty URIs to the subscribed services.

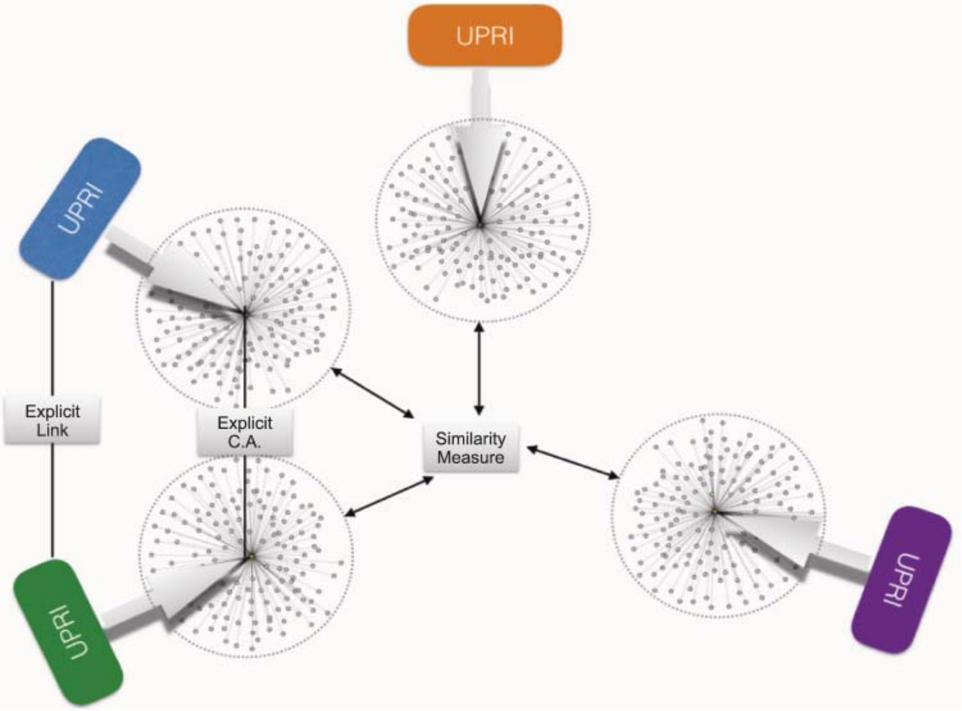


Figure 10. Three ways in which Knowlets can be used to connect dispersed digital objects.

[9] See for instance: https://www.dtls.nl/fair-data/find-fair-data-tools/.

This allows a Web of associated concepts/objects and physical things, including all research objects without the need to explicitly link them ontologically. However, this proposed Web Of Knowlets (WOK) now allows minimally two additional levels of association beyond explicit (ontology-type) links between UPRIs. First, explicit cardinal assertions shared between two never-linked Knowlets can infer a meaningful association between the central concepts in Knowlets. Second, similarity measures (of any kind) over many concepts can reveal a level of conceptual similarity between non-explicitly linked Knowlets (Figure 11).

In blockchain jargon, each Knowlet is also a digital asset and its trusty URI might become part of a chain that can record the ledger of what happened to the Knowlet in time and virtual space. This would be a very interesting new line of research to consider.

Knowlets can change in terms of their conceptual content, which is a transaction that creates a new block with a new hashed UPRI (trusty URI). This enables a massive distributed nodes/miner environment where Knowlets can be traced in various ways. Conceptual drift of Knowlets describing individual abstract concepts such as a disease, a gene, a chemical or a city can be tracked over time and earlier versions can be reproduced to, for instance, visualize conceptual drift, or how subgroups view a concept (for instance patients *versus* medical professionals) or how different religious groups perceive the concept God based on the concepts they associate with it in their formal literature. Knowlets that represent things in the IoT sense can also change and anyone can track what happened to that thing over time. If every nanopublication is treated as a unique meaningful and citable claim, each of those nanopublications can have its own original author, the first person or machine to claim this association in an assertion. The owner can thus also add the nanopublication to the user defined metadata of a data set or a workflow describing FDP (for instance via a CEDAR template) and determine in a smart contract who can do what with the nanopublication. This would also create an automated citation record for the nanopublication (or rather a record of its actual reuse). If a cardinal assertion emerges, and it becomes supported by multiple, identical nanopublications, the cardinal assertion could have its own trusted URI, which again will change when more nanopublications emerge in the IFDS supporting (or commenting on) that cardinal assertion. This might include a contesting of an earlier claim (more and more people assert that they contest this earlier claim), but also assertions that for instance warn about mistakes in a given data set or bugs in a workflow. Each of the players that add an assertion to the IFDS will therefore be able to time-date-stamp that assertion and follow it in a controlled way. It will however always be clear who was the first who made a substantiated claim (related to evidence via its provenance) in the IFDS.

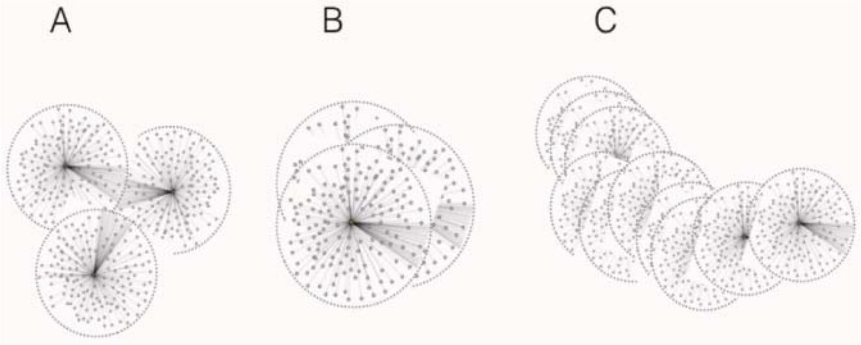


Figure 11. A: Concepts, physical objects or things of different semantic types (and thus also intrinsically meaningless unique, persistent and resolvable identifiers (UPRIs)) can cluster based on contextual similarity without ever being explicitly connected (drug might treat disease). This already works in the Euretos Knowledge Platform for concepts in biomedicine but could also be applied to, for instance, clustering of similar samples and other physical objects, such as specimens in natural history museums. B: Nearly identical concepts that are nevertheless in certain circumstances to be seen as distinct, will automatically cluster as one if the resolution of search or matching is lowered, while they will separate out when the resolution is made higher. Examples are: The same gene in three different species, glucose in solution, and glucose in crystal form, or two thermostats of the same provider, the same batch but located in two different rooms or houses. In all cases, only a few object nodes in their respective Knowlets will differ and therefore they look (correctly) almost identical, but if the right specifications are entered (species matters), (location matters) or if we zoom in enough, they separate. C: Conceptual and semantic drift occur: The meaning of a concept may slightly vary over time, tubes may be moved or databases may be updated, workflows may be versioned, etc. The Knowlet structure has a very powerful ability to capture and record these changes over time. In fact the Knowlet of the different versions of the (time-restricted) meaning of the abstract concept, the physical object or the version of a workflow will change with the changes in the (digital) object it denotes. As Knowlets are intrinsically time-date stamped and their UPRIs may effectively be trusty URIs that automatically change (but stay explicitly linked with the trusty URIs of previous version with time data stamping) over time, and versioning of abstract concepts, workflows, data sets, databases and physical objects can be handled in a scalable way.

As Knowlets are dynamic collections of cardinal assertions about the central subject (concept), many different Knowlets could co-exist (or being created on the fly) describing an abstract concept or a thing in the IFDS. For instance, the Knowlet of a disease could be filtered on predicates, sources (patient/medical professional only), or there could be various perspectives on a particular topic, database or workflow, all pointing to the same object. Obviously, each of these, upon storage in a FAIR data point would create a new trusty URI based on the hash of its content and thus by design have a UPRI. This would enable even payment (of alternative kudos such as scientific credit points, or even cryptocurrencies) on individual nanopublications, cardinal assertions, collections (paper representations even) up to entire databases and trace the composing elements in the collection back to its original owners for proper credit. Most importantly, this would support a completely distributed and non-centrally supervised UPRI creation and blockchain system for data, information and knowledge.

7. DISCUSSION

Why this extra layer? It is not really an extra layer as the Knowlet (a graph) can simply be seen as a FAIR metadata container for the digital object, the physical object or the concept it represents and denotes. A Knowlet is an intrinsically machine-readable graph that defines the denoted digital object/concept or physical object *in its context*, which gives multiple extra options to resolve conceptual drift and near-sameness. The UPRIs of the Knowlets and the digital objects they denote (which have their own UPRI as well) can be routed and resolved via the routing layer at center of the propeller image (Figure 2). As in our reasoning here, metadata Knowlets represent workflows, data sets, articles, and other research objects distributed in the IFDS matching services that can automatically connect data sets with other data sets and with relevant workflows without explicit connections, and conceivably even start automatic distributed analytics. In order to enable the Internet of FAIR Data and Services to be ultimately scalable, a lot of hurdles need to be overcome. Many of them are rather social than technical, but also some technical issues need to be solved by the international communities. In its recent roadmap document for the European Open Science Cloud, the European Commission has reinforced the foundational role of the FAIR principles and has pointed out a series of approaches and initiatives like the Research Data Alliance, collaborating, domain specific research infrastructures and GO FAIR as community-driven driving forces to make this a reality [20].

8. OPEN QUESTIONS

Will a UPRI/Digital Object approach be ultimately scalable to the size we need to allow a global Internet of FAIR Data and Services? Does the proposed nanopublication and Knowlet approach imply a distributing authority for either prefixes, suffixes or both? Can handles be generated automatically without an unacceptable risk to unintendedly duplicate an existing UPRI? Can handles be of the trusty URI type so that they are non-semantic themselves (at least their suffixes) but they are derived from the content of the digital objects they represent? Does every abstract concept (independent from the digital or physical object it may represent) need a UPRI/Handle? Can UUIDs (automatically generated) play a meaningful role? Even if we do not need a distributing authority of UPRIs, do we still need a central registry of all UPRIs to effectively resolve them to digital objects or even physical objects in the IoT? Can we agree on a http://prefix/uuid and/or http://prefix/Trusty URI URL generation system to create preferred handles for all concepts and allow anyone to make additional UPRIs for the same concept as long as they explicitly map it to the GOFAIR handle? Several implementation networks in GO FAIR and in other initiatives are now addressing these questions and I sincerely hope that a convergence to a minimal DO-type approach will bring us a globally scalable solution to enable FAIR science, without the current discrimination against machines.

REFERENCES

[1] B. Mons. Which gene did you mean? BMC Bioinformatics 6(1)(2005),1-4. doi: 10.1186/1471-2105-6-142.

[2] B.M. Good, & M. Wilkinson. The life sciences semantic Web is full of creeps! Briefings in Bioinformatics 7(3) (2006), 275-286. doi:10.1093/bib/bbl025.

[3] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, … & Barend Mons. The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3 (2016), 160018. doi:10.1038/sdata.2016.18.

[4] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L.O.B. da Silva Santos, & M.D. Wilkinson. Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. Information Services & Use 37(1) (2017), 49-56. doi: 10.3233/ISU-170824.

[5] A. Morin, J. Urban, & P. Sliz. A quick guide to software licensing for the scientist-programmer. PLoS Computational Biology 8(7) (2012), e1002598. doi: 10.1371/journal.pcbi.1002598.

[6] The European Commisison High Level Expert Group report: Realising the European Open Science Cloud. Available at: https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud-hleg.

[7] P. Wittenburg, & G. Strawn. Common patterns in revolutionary infrastructures and data. Available at: https://b2share.eudat.eu/records/4e8ac36c0dd343da81fd9e83e72805a0.

[8] M.D. Wilkinson, S.A. Sansone, E. Schultes, P. Doorn, L.O.B. da Silva Santos, & M. Dumontier. A design framework and exemplar metrics for FAIRness. BioRxiv preprint. doi: 10.1101/225490.

[9] B. Mons, M. Ashburner, C. Chichester, E. van Mulligen, M. Weeber, J. den Dunnen, G.J. van Ommen, … & A. Bairoch. Calling on a million minds for community annotation in WikiProteins. Genome Biology 9(2008), R89. doi: 10.1186/gb-2008-9-5-r89.

[10] I.F. Fokkema, J.T. den Dunnen, & P.E. Taschner. LOVD: Easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach. Hum Mutatation 26(2)(2005), 63-68. doi: 10.1002/humu.20201.

[11] Z. Tatum, M.Roos, A.P. Gibson, P.E.M. Taschner, M. Thompson, E.A. Schultes, & J.F.J. Laros. Preserving sequence annotations across reference sequences. Journal of Biomedical Semantics 5(Suppl 1)(2014), S6. doi: 10.1186/2041-1480-5-S1-S6.

[12] B. Mons. Data stewardship for open science: Implementing FAIR principles. Boca Raton: CRC Press, 2018. isbn: 9780815348184.

[13] B. Mons, & J. Velterop. Nano-publication in the e-science era. In: Workshop on Semantic Web Applications in Scientific Discourse, 2009. Available at: https://www.mendeley.com/research-papers/nanopublication-escience-era/.

[14] P. Groth, A. Gibson, & J. Velterop. The anatomy of a nanopublication. Information Services and Use 30(1-2) (2010), 51-56. doi: 10.3233/ISU-2010-0613.

[15] A. Gibson, J.C.J. van Dam, E.A. Schultes, M. Roos, & B. Mons. Towards computational evaluation of evidence for scientific assertions with nanopublications and cardinal assertions. In: CEUR Workshop Proceedings, 2012, pp. 1–7.

[16] T. Kuhn, & M. Dumontier. Trusty URIs: Verifiable, immutable, and permanent digital artifacts for linked data. Video Journal of Semantic Data Management Abstracts 3(1) (2014). Available at: http://videolectures.net/eswc2014_kuhn_linked_data/.

[17] M. Lizio, J. Harshbarger, H. Shimoji, J. Severin, T. Kasukawa, S. Sahin, I. Abugessaisa, … & the FANTOM consortium. Gateways to the FANTOM5 promoter level mammalian expression atlas. Genome Biology 16(2015), 22. doi: 10.1186/s13059-014-0560-6.

[18] H.H.H.B.M. van Haagen, P.A.C. 't Hoen, A. Botelho Bovo, A. de Morre, E.M. van Mulligen, C. Chichester, J.A. Kors, & M.J. Schuemie. Novel protein-protein interactions inferred from literature context. PLoS One 4(11)(2009), e7894. doi: 10.1371/journal.pone.0007894.

[19] H.H.H.B.M. van Haagen, P.A.C. 't Hoen, A. de Morre, W.M.C. van Roon-Mom, D.J.M. Peters, M. Roos, B. Mons, & M.J. Schuemie. In silico discovery and experimental validation of new proteinprotein interactions. Proteomics 11(5)(2011), 843853. doi: 10.1002/pmic.201000398.

[20] Implementation roadmap for the European Open Science Cloud. Available at: https://ec.europa.eu/research/openscience.

AUTHOR BIOGRAPHY

Barend Mons is a Professor of BioSemantics at the Human Genetics Department of Leiden University Medical Centre and founder of the BioSemantics group. He was elected CODATA President in 2018. Next to his leading role in the research of the group, Barend plays a leading role in the international development of "data stewardship" for biomedical data. For instance, he was head-of-node of ELIXIR-NL at the Dutch Techcentre for Life Sciences (until 2015), is Integrator Life Sciences at the Netherlands eScience Centre, and board member of the Leiden Centre of Data Science. In 2014, Barend initiated the FAIR data initiative and in 2015, he was appointed Chair of the European Commission's High Level Expert Group for the "European Open Science Cloud", from which he retired by the end of 2016. Presently, Barend is co-leading the GO FAIR initiative, an initiative to kick start developments towards the Internet of FAIR data and services, which will also contribute to the implementation of components of the European Open Science Cloud. The focus of the contribution of the BioSemantics group is on developing an interoperability backbone for biomedical applications in general and rare disease in particular.